4 Discussion

The segment duration patterns produced by the two speakers are not surprising. Starting with vowel duration, the phonological vowel length is a well established and well known property of Swedish, both as an important feature of Swedish pronunciation, and a way of accounting for the double consonant spelling. As seen in Figure 1, both speakers realize long and short vowel allophones quite similarly. The Swedish speaker, as shown in Figure 1, demonstrates in addition a substantial prolonging of the /p/ segment after short vowel, which the Polish speaker does not. The Polish speaker reports having encountered rules for vowel length as well as consonant length while studying Swedish, implying that mere ignorance does not account for his lack of complementary long consonant. Literature in phonetics, e.g. Ladefoged & Maddieson (1996), gives the impression that phonological vowel length is utilized by a greater number of the world's languages than is consonant length. This suggests that phonological consonant length is a universally more marked feature than is vowel length, and hence more difficult to acquire.

The somewhat greater difference between long and short vowel allophone, demonstrated by the Polish speaker, can be interpreted as a compensation for the lack of complementary consonant length, which is demonstrated to serve as a complementary cue for the listener, when segment durations are in the borderland between /V:C/ and /VC:/ (Thorén 2005).

The between-speaker difference is not surprising, since the phonological quantity in Swedish is a predominant phonetic feature, and can be expected to influence the temporal organization of the native Swede's speech from early age. The Polish speaker came to Sweden as an adult and has acquired one important temporal feature, but his overall temporal organization may still bear strong traces of the system constraints, concerning the duration of segments.

The differences in lip and mandible movements between the speakers could be interpreted as follows: Both speakers produce a higher F1 for short [a] than for long [ɑ:] (e.g. Fant 1959), which typically correlates with lower tongue and mandible. The Polish speaker however, shows a clearly greater jaw and lip opening for long [ɑ:] than for short [a], which suggests that the Polish speaker has a compensatory tongue height in [ɑ:], to maintain correct spectral quality. The greater mandible excursion in [ɑ:] can not be the result of an articulatory goal for this vowel, but could possibly be interpreted as an inverse "Extent of Movement Hypothesis" (Fischer-Jörgensen 1964), letting the mandible make a greater excursion owing to the opportunity offered by the long duration of the [ɑ:].

References

Elert, C-C., 1964. *Phonologic studies of Swedish Quantity.* Uppsala: Almqvist & Wiksell.

Fant, G., 1959. Acoustic analysis and synthesis of speech with application to Swedish. *Ericsson Technics. No. 15*, 3-108.

Fischer-Jörgensen, E., 1964. Sound Duration and Place of articulation. *Zeitschrift für Sprachwissenschaft und Kommunikationsforschung 17*, 175-207.

Ladefoged, P. & I. Maddieson, 1996. *The sounds of the World's Languages.* Oxford: Blackwell publishers.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle & Marchal (eds.), *Speech production and speech modeling*. Dordrecht: Kluwer, 403-439.

Thorén, B., 2005. The postvocalic consonant as a complementary cue to the quantity distinction in Swedish – a revisit. *Proceedings from FONETIK 2005,* Göteborg University, 115-118.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 137
Working Papers 52 (2006), 137–140

Cross-modal Interactions in Visual as Opposed to Auditory Perception of Vowels

Hartmut Traunmüller

Department of Linguistics, Stockholm University
hartmut@ling.su.se

Abstract

This paper describes two perception experiments with vowels in monosyllabic utterances presented auditorily, visually and bimodally with incongruent cues to openness and/or roundedness. In the first, the subjects had to tell what they heard; in the second what they saw. The results show that the same stimuli evoke a visual percept that may be influenced by audition and may be different from the auditory percept that may be influenced by vision. In both cases, the strength of the influence of the unattended modality showed between-feature variation reflecting the reliability of the information.

1 Introduction

Nearly all research on cross-modal interactions in speech perception has been focused on the influence an optic signal may have on auditory perception. In modeling audiovisual integration, it is common to assume three functional components: (1) auditory analysis, (2) visual analysis and (3) audiovisual integration that is assumed to produce an 'amodal' phonetic output. Although details differ (Massaro, 1996; Robert-Ribes et al., 1996; Massaro & Stork, 1998), the output was commonly identified with what the subjects *heard,* not having been asked about what they *saw.* This experimenter behavior suggests the amodal representations of phonetic units (concepts), which can be assumed to exist in the minds of people, to be closely associated with auditory perception. The *seen* remains outside the scope of these models unless it agrees with the *heard.*

The present experiments were done in order to answer the question of whether a visual percept that may be influenced by audition can be distinguished from the auditory percept that may be influenced by vision and whether the strength of such an influence is feature-specific. Previous investigations (Robert-Ribes et al., 1998; Traunmüller & Öhrström, in press) demonstrated such feature-specificity in the influence of optic information on the auditory perception of vowels: the influence was strongest for roundedness, for which the non-attended visual modality offered more reliable cues than the attended auditory modality. In analogy, we could expect a much stronger influence of non-attended acoustic information on the visual perception of vowel height or "openness" as compared with roundedness.

2 Method

2.1 Speakers and speech material

For the two experiments performed, a subset of the video recordings made for a previous experiment (Traunmüller & Öhrström, in press) was used. It consisted of the 6 incongruent auditory-visual combinations of the nonsense syllables /giːg/, /gyːg/ and /geːg/ produced by each one of 2 male and 2 female speakers of Swedish. Synchronization had been based on the release burst of the first consonant. In Exp. 1, each auditory stimulus was also presented alone and in Exp. 2 each visual stimulus instead.

2.2 Perceivers

14 subjects believed to have normal hearing and vision (6 male, aged 20 to 60 years, and 8 female, aged 20 to 59 years) served as perceivers. All were native speakers of Swedish pursuing studies or research at the Department of Linguistics.

2.3 Procedure

The subjects wore headphones AKG K25 and were seated with their faces at an arm's length from a computer screen. Each one of the 9 stimuli per speaker was presented twice in random order, using Windows Media Player. The height of the faces, shown in the left half of the screen, was roughly 12 cm. In Exp. 1, the subjects were instructed to look at the speaker when shown, but to tell which vowel they *heard.* In Exp. 2, they were instructed to keep the headphones on, but to tell which vowel they *saw.* Stimulus presentation was controlled individually by the subjects, who were allowed to repeat each stimulus as often as they wished. They gave their responses by clicking on orthographic symbols of the 9 Swedish vowels arranged in the right half of the screen in manner of an IPA-chart. They were told to expect a rather skewed vowel distribution. Prior to each experiment proper, three stimuli were presented for familiarization. The two experiments lasted for grossly 30 minutes together.

3 Results

The pooled results of each one of the two experiments are shown in Tables 1 and 2. It can be noticed that in Exp. 1, where subjects had to report the vowels they heard, openness was almost always perceived in agreement with the speaker's intention (99.2%) even when conflicting optic information was presented. Roundedness was perceived much less reliably: 14.7% errors when no face was presented. Many of these errors were evoked by one particular speaker. In cases of incongruent information, roundedness was predominantly perceived in agreement with the optic rather than the acoustic stimulus.

The picture that emerged from Exp. 2, where subjects had to report the vowels they saw by lipreading, was the reverse: Presence vs. absence of roundedness was perceived correctly to 98.4% and the rounded vowels were only in 5.4% of the cases mistaken for inrounded (labialized), while openness was perceived quite unreliably, with 28.3% errors when no acoustic signal was presented. (One of the speakers elicited substantially fewer errors.) In cases of incongruent information, openness was often perceived in agreement with the acoustic rather than the optic stimulus, but the cross-modal influence was not as strong in lipreading (Exp. 2) as in listening (Exp. 1). This can be seen more immediately in Table 3, which shows the result of linear regression analyses in which the numerical mean of the perceived openness and roundedness of each stimulus were taken as the dependent variables.

While the overall close to perfect performance in auditory perception of openness and in visual perception of roundedness by all speakers does not allow any possible between-perceiver differences to show up, such differences emerged, not unexpectedly, in auditory roundedness perception and in visual openness perception (see Table 4). In auditory roundedness perception, the case-wise reliance on vision varied between 31% and 100%. In visual openness perception, the case-wise reliance on audition varied between 28% and 97%. Despite the similarity in range, there was no significant correlation between these two variables (r^2=0.04, p=0.5), nor was there any significant gender difference in visual perception (p>0.4). This means that the between-subject variation cannot be explained as due to a subject specific (or gender specific) general disposition towards cross-modal integration. In auditory perception, women showed a greater influence of vision, but the gender difference failed to attain significance (two-tailed t-test, equal variance not assumed: p=0.15), and age was never a significant factor.

Table 1. Confusion matrix for auditory perception. Stimuli: intended vowels presented acoustically and optically. Responses: perceived vowels (letters). Correct vowel; *incorrect openness*; roundedness incorrect but agrees with optic stimulus. Boldface: majority response.

Stimuli		*Responses*									
Sound	*Face*	*i*	*y*	*u*	*o*	*e*	*ö*	*å*	*ä*	*a*	*
i		102	9								1
y		17	94								1
e						88	23				1
i	e	109	3								
i	y	27	**85**								
y	i	**85**	26			*1*					
y	e	**84**	22			*6*					
e	i					104	8				
e	y			*1*		38	**73**				

Table 2. Confusion matrix for visual perception. As in Table 1, but *incorrect roundedness*; openness incorrect but agrees with acoustic stimulus.

Stimuli		*Responses*									
Sound	*Face*	*i*	*y*	*u*	*o*	*e*	*ö*	*å*	*ä*	*a*	*
	i	56				54				2	
	y	*1*	73	7	3		28				
	e	9				101			2		
e	i	16				**91**	*4*				1
y	i	77	2			31	*1*		1		
i	y	*1*	95	5			11				
e	y		62	3		*1*	46				
i	e	**65**				46	*1*				
y	e	52	2			57				*1*	

Table 3. Weights of auditory and visual cues in the perception of openness and roundedness.

	Auditory cues	*Visual cues*
Heard openness	1.00	0.03
Heard roundedness	0.28	0.68
Seen openness	0.45	0.52
Seen roundedness	0.00	0.97

Table 4. Auditory roundedness and visual openness perception by subject (age in years, sex). Percentage of responses in agreement with the acoustic (Aud) or optic (Vis) stimulus in cases of incongruent information (n=32). Incorrect but agrees with the unattended modality.

Subject	*Age*	*34*	*27*	*30*	*20*	*41*	*34*	*21*	*23*	*60*	*27*	*20*	*23*	*25*	*59*
	Sex	*m*	*m*	*f*	*m*	*m*	*f*	*f*	*f*	*m*	*f*	*m*	*f*	*f*	*f*
Roundedness	Aud	69	59	44	41	31	34	25	22	9	6	6	3	3	0
of vowels *heard*	Vis	31	41	**50**	**59**	**63**	**66**	**75**	**78**	**91**	**94**	**94**	**97**	**97**	**100**
Openness	Aud	**59**	47	28	34	**72**	**69**	**97**	**69**	**63**	**72**	**50**	**75**	28	**59**
of vowels *seen*	Vis	41	53	72	66	28	31	3	31	38	28	47	25	72	41

The patterns of confusions can be modelled by weighted summation of the response probabilities for each vowel in the attended modality [listening (A), lipreading (V)] and a Bayesian auditory-visual integration (AV). For the pooled data, linear regression on this basis gives response probabilities P and determination coefficients r^2 as follows:

$P_{heard} = 0.01 + 0.26\ A + 0.71\ AV\ (r^2 = 0.98)$ and $P_{seen} = -0.00 + 0.57\ V + 0.45\ AV\ (r^2 = 0.94)$.

4 Discussion

As for auditory perception with and without conflicting visual cues and for visual perception alone (lipreading), the patterns of confusion observed here agree closely with those obtained previously (Traunmüller & Öhrström). Now, the novel results obtained in visual perception with conflicting auditory cues demonstrate that a visual percept that may be influenced by audition has to be distinguished from the auditory percept that may be influenced by vision and that the strength of the cross-modal influence is feature-specific in each case.

Based on confusion patterns in consonant perception, it has been claimed that humans behave in accordance with Bayes' theorem (Massaro & Stork, 1998), which allows predicting bimodal response probabilities by multiplicative integration of the unimodal probabilities. Although some of our subjects behaved in agreement with this hypothesis in reporting what they *heard*, the behaviour of most subjects refutes the general validity of this claim, since it shows a substantial additive influence of the auditory sensation. When reporting what they *saw*, all subjects except one showed a substantial additive influence of the visual sensation.

Given the unimodal data included in Tables 1 and 2, Bayesian integration lends prominence to audition in the perception of openness and to vision in roundedness. The data make it clear that an ideal perceiver should rely on audition in the perception of openness, as all subjects did in their auditory judgments, and combine this with the roundedness sensed by vision, since this is more reliable when the speaker's face is clearly visible. Four female and two male subjects behaved in this way to more than 90% in reporting what they *heard* but only one other female subject in reporting what she *saw*.

The results can be understood as reflecting a weighted summation of sensory cues for features such as openness and roundedness, whereby the weight attached reflects the feature-specific reliability of the information received by each sensory modality (cf. Table 3). The between-perceiver variation then reflects differences in the estimation of this reliability.

Acknowledgements

This investigation has been supported by grant 2004-2345 from the Swedish Research Council. I am grateful to Niklas Öhrström for the recordings and for discussion of the text.

References

Massaro, D., 1996. Bimodal speech perception: a progress report. In D.G. Stork & M.E.Hennecke (eds.), *Speechreading by Humans and Machines*. Berlin: Springer, 80-101.

Massaro, D.W. & D.G. Stork, 1998. Speech recognition and sensory integration. *American Scientist 86*, 236-244.

Robert-Ribes, J., M. Piquemal, J-L. Schwartz & P. Escudier, 1996. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In D.G. Stork & M.E.Hennecke (eds.), *Speechreading by Humans and Machines*. Berlin: Springer, 193-210.

Robert-Ribes, J., J-L. Schwartz, T. Lallouache & P. Escudier, 1998. Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America 103*, 3677-3689.

Traunmüller, H. & N. Öhrström, in press. Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 141
Working Papers 52 (2006), 141–144

Knowledge-light Letter-to-Sound Conversion for Swedish with FST and TBL

Marcus Uneson

Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
marcus.uneson@ling.lu.se

Abstract

This paper describes some exploratory attempts to apply a combination of finite state transducers (FST) and transformation-based learning (TBL, Brill 1992) to the problem of letter-to-sound (LTS) conversion for Swedish. Following Bouma (2000) for Dutch, we employ FST for segmentation of the textual input into groups of letters and a first transcription stage; we feed the output of this step into a TBL system. With this setup, we reach 96.2% correctly transcribed segments with rather restricted means (a small set of hand-crafted rules for the FST stage; a set of 12 templates and a training set of 30kw for the TBL stage).

Observing that quantity is the major error source and that compound morpheme boundaries can be useful for inferring quantity, we exploratively add good precision-low recall compound splitting based on graphotactic constraints. With this simple-minded method, targeting only a subset of the compounds, performance improves to 96.9%.

1 Introduction

A text-to-speech (TTS) system which takes unrestricted text as input will need some strategy for assigning pronunciations to unknown words, typically achieved by a set of letter-to-sound (LTS) rules. Such rules may also help in reducing lexicon size, permitting the deletion of entries whose pronunciation can be correctly predicted from rules alone. Outside the TTS domain, LTS rules may be employed for instance in spelling correction, and automatically induced rules may be interesting for reading research.

Building LTS rules by hand from scratch is easy for some languages (e.g., Finnish, Turkish), but turns out prohibitively laborious in most cases. Data-driven methods include artificial neural networks, decision trees, finite-state methods, hidden Markov models, transformation-based learning and analogy-based reasoning (sometimes in combination). Attempts at fully automatic, data-driven LTS for Swedish include Frid (2003), who reaches 96.9 % correct transcriptions on segment level with a 42000-node decision tree.

2 The present study

The present study tries a knowledge-light approach to LTS conversion, first applied by Bouma (2000) on Dutch, which combines a manually specified segmentation step (by finite-state transducers, FST) and an error-driven machine learning technique (transformation-based learning, TBL). One might think of the first step as redefining the alphabet size, by introducing new, combined letters, and the second as automatic induction of reading rules on that (redefined) alphabet, ordered in sequence of relevance.

For training and evaluation, we used disjoint subsets of a fully morphologically expanded form of Hedelin et al. (1987). The expanded lexicon holds about 770k words (including